Creating Generative Lexicon form Dictionaries: Tamil Experience

S. Rajendran
Associate Professor
Department of Linguistics
Tamil University
Thanjavur 613 005
Email: raj_ushush@yahoo.com

Abstract

Pustejovsky (1995) characterizes a generative lexicon as a computational system involving at least the following levels of representation: argument structure, event structure, qualia structure and lexical inheritance structure. It is worth exploring the possibility of taping all these information for building a generative lexicon from MRDs and other dictionaries. Machine Readable dictionaries (MRDs) contain extensive linguistic information on lexical semantics. In dictionaries the lexicographers define words in terms of other words, which themselves have definitions elsewhere in the dictionary. The computational linguists can unpack the meanings that have between encoded in dictionaries by lexicographers by converting them into a computationally tractable form. By creating a lexical database in this way the generative lexicon proposed by Pustejovsky can be built.

1. Introduction

There is rising interest in computational analysis of Machine Readable Dictionaries (MRDs) as a lexical resource for various purposes in the recent years (Wilsk et al, 1996:161-181). Existing dictionaries are obvious place to look for extensive linguistic information on lexical semantics. Exploiting MRDs for semantic analysis is a clear-cut effort, which aims at harvesting the work-years of effort by professional lexicographers. It is a well-known fact that a wealth of implicit information lies within dictionaries. This lexical information can be made explicit for use by programs. In dictionaries the lexicographers define words in terms of other words, which themselves have definitions elsewhere in the dictionary. The computational linguists have to unpack the meanings that have between encoded in dictionaries by lexicographers by converting them into a computationally tractable form. The motivation for creating lexical database in this way is to take advantage of the extensive lexicographic work that goes into creating the dictionary in the first place. It is proposed here to explore the possibility of taping all the information required for building a generative lexicon from MRDs and other dictionaries.

2. Generative Lexicon

The generative lexicon (shortly GL) presents a novel and exciting theory of lexical semantics that addresses the problem of the "multiplicity of word meaning"- that is, how we are able to give an infinite number of senses to words with finite means. As the first formally elaborated theory of generative approach to word meaning, it lays the foundation for an implemented computational treatment of word meaning that connects explicitly to a compositional semantics. In contrast to static view of word meaning (where each is word is

characterized by a predetermined number of word senses) that imposes a tremendous bottleneck on the performance capability of any natural language processing, Pustejovsky proposes that the lexicon becomes an active and central component in the linguistic description. The essence of his theory is that the lexicon functions generatively, first by providing a rice and expressive vocabulary for characterizing lexical information; then by developing a frame work for manipulating fine-grained distinctions in word descriptions; and finally, by formalizing a set of mechanisms for specialized composition of aspects of such description of words, as they occur in context, extended and novel senses are generated.

According to Pustejovsky (1995:5) the most pressing problem in lexical semantics are the following:

1. Explaining the polymorphic nature of language
2. Characterizing the semanticality of natural language utterances
3. Capturing the creative use of words in novel contexts.
4. Developing a richer, co-compositional semantic representation

Pustejovsky (2001:56) characterize a generative lexicon as a computational system involving at least the following levels of representation

1. ARGUMENT STRUCTURE: Specification of number and type of logical arguments
2. EVENT STRUCUTE: Definition of the event type of an expression and its subeventual structure
3. QUALIA STRUCTURE: A structural differentiation of the predicative force for a lexical item
4. LEXICAL INHERITANCE STRUCTURE: Identification of how a lexical structure is related to other structures in this type of lattice

Pustejovsky (2001:56) assumes that word meaning is structured on the basis of four generative factors, or qualia roles, that capture how humans understand objects and relations in the world and provide the minimal explanation for the linguistic behaviour of lexical items.

CONSTITUTIVE: the relation between an object and its constituent parts
FORMAL: the basic category that distinguishes the object within a larger domain
TELIC: the object's purpose and function
AGENTIVE: factors involved in the object's origin or "coming into being."

The qualia structure is the core of the generative properties of the lexicon, because it provides a general strategy for creating increasingly specific concepts with conjunctive properties. A simple schematic description of a lexical item, α, using this representation is shown below:

α
ARGSTR = ARG1=x
...

QUALIA = CONST = what x is made of
FORMAL = what x is
TELIC = function of x

AGENTIVE = how x come into being

The lexical structure for *book* as an object can then be represented as follows:

book		
	ARG1=	y:information
ARGSTR =	ARG2=	x:phys_obj
		information.phy_obj
	FORM =	holds (x,y)
QUALIA =	TELIC =	read (e,w,x.y)
	AGENT =	write (e', v, x, y)

3. Extraction of semantic information form dictionaries

Much of the information needed for our purpose can be found in a dictionary. The problem, of course, is that none of this comes without great effort. Every dictionary is printed in a different format, and every format is designed to be both information-bearing and pleasing to the eye. This combination makes it very difficult to remove or reformat the strings and symbols in the dictionary automatically, to decide, by program, what is intended by the lexicographer, what to keep, and what to leave out. Extraction of semantic information from dictionaries entails certain assumptions about the extent of knowledge in a dictionary, about where that knowledge is located, and how that knowledge can be extracted from the language of dictionary definitions. These assumptions are methodological assumptions because they underlie the decisions made in choosing one method for semantic analysis rather than another. These assumptions are about sufficiency, extricability, and bootstrapping (Wilks et al 1990, 1996:141-142).

Sufficiency addresses the issue of whether a dictionary is a strong enough knowledge base for Tamil, specially as regards linguistic knowledge and, above all, the knowledge of the real world needed for subsequent text analysis. Different positions have been taken toward sufficiency within computational lexicography. Some researchers believe that there is not enough knowledge in dictionaries in principle; in other words, that certain specific semantic information is not available anywhere in a dictionary and hence must be derived from another, outside, source. Other researchers believe that dictionaries may indeed contain sufficient knowledge, although that knowledge may be implicit, because that knowledge can be made explicit by using information from entries in other parts of a dictionary.

Extricability is concerned with whether it is possible to specify a set of computational procedures that operate on an MRD and extract, through their operation alone and without any human intervention, general and reliable semantic information on a large scale, and in a general format suitable for, though independent of, a range of subsequent NLP tasks.

Bootstrapping refers to a process of collecting the initial information that is required by a set of computational procedures that are able to extract semantic information from the sense definitions in an MRD. The initial information needed is commonly linguistic, normally syntactic and case information, which is used during the processing of dictionary sense definitions into an underlying representation form which semantic information is then extracted. Boost strapping methods can be internal or external. Internal bootstrapping

methods obtain the initial information needed for their procedures from the dictionary itself and use procedures to extract that information. By contrast, external bootstrapping methods obtain the initial information for their procedures by some method other than the use of those procedures themselves. The initial information may be from some source external to the dictionary or may be in the dictionary but impossible to extract without the use of the very same information. There are differences of opinion in computational lexicology regarding extricability and bootstrapping.

The straight forward approach to creating an MTD is to segment the dictionary into a collection of records, one for each word sense, composed of a set of labeled constituent fields, where each filed is roughly equivalent to a single meaningful element in a sense definition. The procedure makes the contents of the dictionary explicit, and is the so-called lexical database approach for forming an MTD. The more difficult approach is to process the labeled constituent fields in some way, in order to retrieve the implicit knowledge that lies buried there. This is the semantic approach to creating an MTD.

Pustejovsky defines the semantics of a lexical item as a structure involving different components. One of these is qualia structure, which is rich and structured representation of the rational force of a lexical item. What is peculiar about GL is that qualia permits a much richer description of meaning than either a simple decompositional view or a purely relational approach to word meaning allow. That is, it expresses different/orthogonal aspects of word meaning instead of a one-dimensional inheritance (even multiple), which can only capture standard hyponymy/hypernymy relations. The adequacy of qualia relations for capturing key aspect of word meaning becomes apparent when consulting dictionary definitions. The elements of meaning easily map on dimension(s) expressed via qualia roles. Furthermore, these relations become particularly crucial for those sense definitions which have an underspecified genus term. Not all dimension of meaning are always explicitly expressed in the definition of a lexical item. Some of them are inherited by virtue of its membership to a semantic type. Although qualia relations easily emerge form dictionary definition, the formal expression of a specific value for a quale is sometimes quite problematic.

4. Noun Semantics

Dictionary definitions of nouns are normally written in such a way that one can identify for each headword, a "genus term", and these are related via an IS-A relation. Information following the genus term, the differentia, serves to differentiate the headword from other headwords with the same genus. Now let us examine the meanings of certain lexical items given in *kiriyaavin taRkaalattamiz akaraati* (shortly KTTA, which in machine readable form) and explore how much of information we can gather for constructing qualia structure: (The English glosses are given to facilitate understanding.)

> *puttakam* 'book' is defined in KTTA as follows:
> *puttakam pe*. 1. *paTippataRku eeRRavaliaiyil aTTaipooTTu iNaitta acciTTa taaLkaLin tokuppu* 'printed pages tied together by binding into a collection to enable reading.'2. *ezhutuvataRku eeRRavakaiyil veRRuttaazkaLai aTTai pooTTu iNaitta tokupp*pu 'plain paper tied together into a collection to enable writing'.

tokuppu 'collection' is the genus (i.e. formal role) of *puttakam*, which can be easily identified being the head noun of the clause. *aTTaipooTTu iNaitta* and *accaTitta taaLkaL* 'printed

papers' are attributes (constitutive role) which helps to distinguish *puttakam* 'book' form other printed materials, for example, *accaTitta aRivippu* 'printed notice'. *paTippataRku* 'for reading' stands to denote the function (i.e. telic role) of *puttakam*. From the definition given in KTTA we can arrive at the following qualia structure for *puttakam*.

***puttakam* 'book**

ARGSTR = ARG1= x: phys_obj

CONSTITUTIVE = made_ of: *aTTai* 'cover', *accaTitta taaLkaL* 'printed papers'
FORMAL = *tokuppu* 'collection' (x)
QUALIA = TELIC = *paTi* 'read' (x,z)
AGENTIVE = *accaTi* 'print' (x), *pooTTu iNai* 'bind by': *aTTai* 'cover'

The second meaning stands to denote a *notebook.* The agentive information that *puttakam* is the product of writing is not given in the meaning. Instead it is given as the product of printing. Also the fact that *puttakam* is ambiguous like *book*, denoting information and physical object cannot be had from the definition.

Let us look at the meaning of the *katti* found in KTTA:

katti 'knife' *paTTaiyaana ulookattakaTTin oorangkaLil kuurmai uTaiyataaka uLLa veTTum karuvi* 'a cutting instrument of flat meal plate with sharp edges '

In the definition *karuvi* 'instrument' is the genus term of the headword *katti* and *paTTaiyaana ulookattakaTTin oorangkaLil kuurmai uTaiyataaka uLLa* 'flat meal plate with sharp edges' and *veTTum* 'that which can cut' yields differentia. The noun phrase *paTTaiyaana ulookattakaTTin* 'in band of metal plate' gives made_of constitutive role and the clause *oorangkaLil kuurmai uTaiyataaka uLLa* 'with sharp edges' gives the has_as_ part constitutive role and the *veTTum* gives the telic role. *katti* can be represented by the following qualia structure.

***katti* 'knife'**

ARGSTR = ARG1= x: phys_obj

CONSTITUTIVE = made_of : *paTTaiyaana ulookattakaTu* 'band of metal plate'. has_as_part: *kuurmaiyaana oorangkaL* 'sharp
FORMAL = *karuvi* 'instrument' (x)
QUALIA = TELIC = *veTTu* 'cut' (x,z)
AGENTIVE = ?

That agentive role that *katti* is an artifact is not explicitly spelt in the definition, but it can be inferred form the genus term *karuvi* 'instrument'.

5. Verb Semantics

The qualia structure associated with causative predicates (i.e. TRANSACTIONS) verbs are typically analyzed as involving an initial act or process or followed by a resulting state. These two phrases map directly into the AGENTIVE and FORMAL qualia roles, respectively. The transitive form of the verb *break* is illustrated below (Pustejovsky (1995:80):

break

EVENTSTR= E1 = e1: process
E2 = e2:state
RESTR = <α

QUALIA= FORMAL = broken (e2,y)
AGENTIVE = break_act (e1,x,y)
...

The qualia structure of *ran* and *sleep* are differentiated by the AGENTIVE and FORMAL roles.

run

EVENTSTR = E1 = e1:process

QUALIA = AGENTIVE = run_act (e1,x)
...

sleep

EVENTSTR = E1 = e1:process

QUALIA = FORMAL = sleep(e1,x)

Now let look examine definitions of a few verbs and how far the information fulfill qualia structure. The verb *tangku* 'stay' is described in KTTA as follows:

tangku 'stay' 1: (*kuRippiTTa ndookkattukkaaka oor iTattukkuc cenRu) kuRippiTTa kaalam varai iruttal* 'to go and be at a place for a fixed period of time with a purpose'

Busa et al (1995) represent the lexical structure of *stay* in the following fashion:

stay

type: **simple:location_formal**$_{isa}$

argstr:
arg1: **x: living_entity**
arg2: **y: location**

evenstr:[e1:**e1:state]**

qualia: [formal:**at (e1,x,y)**

The definition in KTTA emphasizes being at a place for certain period. The lexical structure given by Busa et al does not account for time, only location is accounted. The information needed for lexical structure can be had form the definition and the illustrative sentences. KTTA defines *anuppu* 'send' in the following fashion:

> *anuppu* 'send'1: (*oruvarai*) *oor iTattukkup pookacceytal* 'to make a person go to a place'; (*onRai*) *oor iTam cenRaTaiyac ceytal* 'causing a thing to reach a place'
>
> **send**
> type: **unifed:location_formal**$_{\text{result_}}$**agentive**$_{\text{_cause_}}$**telic**$_{\textbf{purp}}$
>
> argustr: arg1 : **x:living_entity**
> arg2 : **y:entity**
> arg3 : **z:entity**
>
> evenstr: e1: **e1:state**
> e2: **e2:process**
> rest: **e1 precedes e2**
>
> qualia: formal **(¬at (e1,x,y)**
> telic: **at (e1,z,y)**
> agentive: **act (e2,x,y)**

6. Polysemy

A proper treatment of the description of the semantics of lexical items in the language should permit us to adequately describe the behaviour of ambiguity as well as the lexical process of selection process in the grammar. Weinreich (1964) distinguishes two types of ambiguity, the first of which he calls contrastive ambiguity. This seen where a lexical item accidentally carries two distinct and unrelated meanings (i.e. homonymy). For example, the lexical item *bank* found in the following sentences exhibit contrastive senses.

> Mary walked along the bank of the river.
> HarborBank is the richest bank in the city.

Complementary ambiguity exhibits a different type of relation between senses. The sense alternation is just one of many alternations that can be described as logical polysemies, where the noun seems to have systematically related senses.

> Count/Mass alternations; lamb
> a. The lamb is running in the field.
> b. John ate lamb for breakfast
>
> Container/Containee alternations; bottle
> a. Mary broke the bottle.
> b. The baby finished the bottle.

The most direct way to account for the polysmies described above is to allow the lexicon to have multiple listing of words, each annotated with a separate meaning or lexical sense. This is certainly the simplest means of encoding sense variation in a lexical item. Pustejovsky (1995:34) calls such a dictionary as a Sense Enumeration Lexicon (SEL), and characterize it is directly as follows:

> A Lexicon L is a Sense Enumeration Lexicon if and only if for every word w in L, having multiple senses s1,...,sn, associated with the word, then the lexical entries expressing these senses are sorted as {ws1,...,Wsn}.

SEL represent the contrastive senses of *bank* in the following fashion.

> **bank1**
> CAT = **count_nou**n
> GENUS = **financial_institution**
>
> **bank2**
> CAT = **count_noun**
> GENUS = **shore**

Also SEL represents the complementary senses of *lamb* in the following fashion.

> lamb1
> CAT = count_noun
> GENUS = meat
>
> lamb2
> CAT= mass_noun
> GENUS = meat

Pustejovsky (1995: 38) restate the SEL's account of contrastive and complementary senses as follows:

> A Lexicon L is a Sense Enumeration Lexicon if and only if for every word w in L, having multiple senses $s_1,...,s_n$ associte with that word, then:
>
> (i) if $s_1,...,s_n$ are contrastive senses, the lexical entries expressing these senses are stored as $w_{s1},...w_{sn.}$
> (ii) if $s_1,...s_n$ are complementary senses, the lexical entry expressing these senses is stored as w{s1,...sn}.
>
> Every ambiguity is either represented by (i) or (ii) above.

Though Pustejovsky points out the advantage of this model of lexical description, he also points out that the SEL model is inadequate for the purpose of linguistic theory. But unfortunately KTTA does not stand to fulfill the requirement of a SEL model completely. For example, KTTA does not enumerate the mass-noun sense of the lexical item *koozi* 'fowl', though it is a possible sense.

koozi pe. (*potuvaaka*) *viiTu, paNNai aakiyavaRRil muTTaikkaakavum iRaiccikkaakavum vaLarkkappaTum atika uyaram paRakkaata p*aRavai 'a type of bird which cannot fly high and which are generally reared in houses, farms, etc. for the sake of eggs and meat'

The lexical item jannal '*window*' is polysemous expressing aperture sense and physical object sense which can be represented as follows:

window1
CAT = count_noun
GENUS = aperture

window2
CAT = count_noun
GENUS = physical_obj

KTTA does not account for the the physical object sense of *jannal* 'window'.

jannal pe. kaTTiTam poonRavaRRiRil kaaRRum veLiccamum varuvataRku eeRRa vakaiyil tiRandtu muuTakkuTiyataaka cuvaril amaikkappaTTirukkum tiRappu. 'an opening in the wall of a building which can be closed or opened with the facility of the light and wind to come inside'

Pustejovsky (1995:39-60) points out the limitations of SELs. KTTA also shows all these limitations being a dictionary and not a generative lexicon.

7. Conclusion

It is true that dictionaries including MRDs represent meanings of lexical item inadequately. They cannot be equated or compared with generative lexicon propounded by Pustejovsky. But it will not be a waste to utilize the information available in the dictionaries to build a generative lexicon. If one has a tool to extract the information automatically that is a welcome thing, but constructing such a tool is a difficult task. To start with we explore the amount of information available in the dictionaries to build a generative lexicon.

Reference

[Busa et al., 1998] F. Busa, N. Calzolari and A. Lenci. Generative Lexicon and the SIMPLE Model: Developing Semantic Resources for NLP." In F. Busa and P. Boulillon (eds). *Language of Word Meaning.* Cambridge: Cambridge University Press.
[Cruse, 1986] D.A. Cruse. *Lexical Semantics*. Cambridge, England: Cambridge University Press.
------- 2000. *Meaning in Language: An Introduction to Semantics and Pragmatics*. Oxford, New York: Oxford University Press.
[Hearst, 1999] M.A. Hearst. Automated Discovery of WordNet Relations. In C. Fellbaum (ed.). *WordNet: An Electronic Lexical Database*. Cambridge, Mass.: MIT Press.
[Pustejovsky, 1995] J. Pustejovksy. *The Generative Lexicon*. Cambridge, Massachusetts: The MIT Press.

------- 2001. Generativity and Explanation in Semantics: A Reply to Fodor and Lepore. In P. Bouillon F. Busa (eds.). *The Language of Word Meaning*. Cambridge: Cambridge University Press.
[Pustejovky and Boguraev, 1993] J. Pustejovky and B. Boguraev. 1993. Lexical Knowledge Representation and Natural Language Processing. In F. Pereira and B. Grosz (eds.). *Natural Language Processing*. Cambridge, Mass.: MIT Press.
------1994. A Richer Chracterization of Dictionary Entries: The Role of Knowledge Representation. In B.T.S. Atkins, and A. Zampolli (eds.). *Computational Approach to Lexicon.* Oxford: Oxford University Press.
[Rajendran, 1978] S. Rajendran. *Syntax and Semantics of Tamil Verbs* (manuscript). Ph.D. Thesis. Poona: University of Poona.
------1983. *Semantics of Tamil Vocabulary*. (Report of the UGC sponsored Postdoctoral Work in manuscript). Poona: Deccan College Post Graduate and Research Institute.
------ 2002. Preliminaries to the Preparation of a Word Net for Tamil. *Language in India* 2:1, www.langugeinindia.com
[Rajendran et al., 2002] S. Rajendran, S. Arulmozi, B. Kumara Shanmugam, S. Baskaran, and S. Thiyagarajan. Tamil WordNet. In *Proceedings of the First International Global WordNet Conference*. Mysore: CIIL, PP: 271-274.
[Ruimy et al., 2001] N. Ruimy, E. Gola and M. Monachini. Lexicography Informs Lexical Semantics: The SIMPLE Experience. In F. Busa, and P. Boulillon (eds). *Language of Word Meaning*. Cambridge: Cambridge University Press.
[Weinreich, 1964] U. Weinreich. Webster's Third: A Critique of Semantics. *International Journal of American Linguistics* 30, PP: 405-409.
[Wierzbica, 1980] A. Wierzbica. *Lingua Mentalis: The Semantics of Natural Language*. New York: Academic Press.
----------1988. *The Semantics of Grammar*. Amsterdam: John Benjamins.
[Wilks et al. 1990] Y.A. Wilks, D.C. Fass, C.M. Guo, J.E. McDonald, T. Plate and B.M. Slator. Providing Machine Tractable dictionary tools. In J. Pustejovksy (ed.). *Semantics and the Lexicon*. Cambridge, Mass.: MIT Press.
[Wilks, et al., 1996] Y.A. Wilks, B.M. Slator and L.M. Guthrie. Electric Words: Dictionaries, Computers, and Meanings. Cambridge, Mass.: MIT Press.